

(19) World Intellectual Property Organization  
International Bureau(43) International Publication Date  
22 November 2001 (22.11.2001)

PCT

(10) International Publication Number  
WO 01/88713 A2

(51) International Patent Classification<sup>7</sup>: G06F 12/00

(21) International Application Number: PCT/US01/15591

(22) International Filing Date: 15 May 2001 (15.05.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:  
60/204,184 15 May 2000 (15.05.2000) US

(71) Applicant: SUN MICROSYSTEMS, INC. [US/US]; 901 San Antonio Road, MS PALI-521, Palo Alto, CA 94303 (US).

(72) Inventors: FLOOD, Christine, H.; 13 Main Street, Westford, MA 01886 (US). AGESEN, Ole; 154 Laurel Drive, Needham, MA 02492 (US). DETLEFS, David, L.; 94 Depot Street, Westford, MA 01886 (US). SHAVIT, Nir, N.;

(153) Address for Correspondence: 153 Upland Road, Cambridge, MA 02140 (US). ZHANG, Xiaolan; 9 Crescent Street, Cambridge, MA 02138 (US).

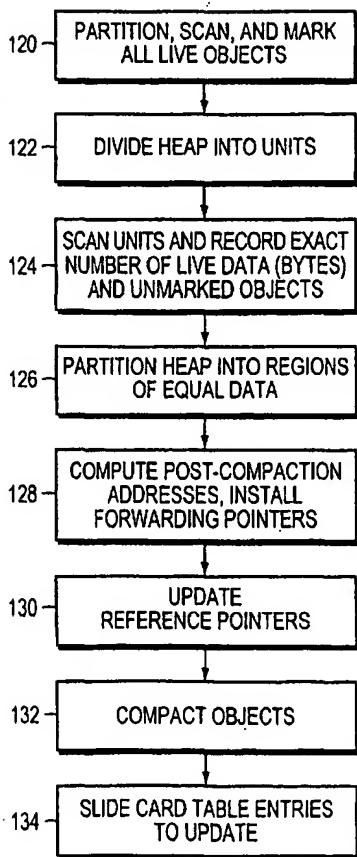
(74) Agents: PAUL, Edwin, H. et al.; Cesari and McKenna, LLP, 88 Black Falcon Avenue, Boston, MA 02210 (US).

(81) Designated States (national): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.

(84) Designated States (regional): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: WORK-STEALING QUEUES FOR PARALLEL GARBAGE COLLECTION



(57) Abstract: A multiprocessor, multi-program, stop-the-world garbage collection program is described. The system initially over partitions the root sources, and then iteratively employs static and dynamic work balancing. Garbage collection threads compete dynamically for the initial partitions. Work stealing double-ended queues, where contention is reduced, are described to provide dynamic load balancing among the threads. Contention is resolved by using atomic instructions. The heap is broken into a young and an old generation where parallel semi-space copying is used to collect the young generation and parallel mark-compacting the old generation. Speed and efficiency of collection is enhanced by use of card tables and linking objects, and overflow conditions are efficiently handled by linking using class pointers. A garbage collection termination employs a global status word.

WO 01/88713 A2



**Published:**

— without international search report and to be republished upon receipt of that report

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

## WORK-STEALING QUEUES FOR PARALLEL GARBAGE COLLECTION

### CROSS REFERENCE TO RELATED APPLICATIONS

This application is related to and claims priority from Provisional Application, 5 Serial No. 60/204,184, filed on June 15, 2000, and entitled "Parallel Garbage Collection for Shared Memory Multiprocessors."

This application is also closely related to commonly assigned U.S. applications of Christine H. Flood et al. entitled: Local Allocation Buffers for Parallel Garbage Collection; Using Atomic Compare-and-Swap operations for Forwarding-Pointer Installation; and Striding-Type Generation Scanning for Parallel Garbage Collection, all 10 co-filed with the present application.

### BACKGROUND OF THE INVENTION

#### *Field of the Invention*

The present invention relates to memory management particularly with the aspect of memory management that has become known as "garbage collection." More 15 particularly the present invention relates to garbage collection in systems having multiple processors sharing memory.

#### *Background Information*

In the field of computer systems, considerable effort has been expended on the 20 task of allocating memory to data objects. For the purposes of this discussion, the term *object* refers to a data structure represented in a computer system's memory. Other terms sometimes used for the same concept are *record* and *structure*. An object may be identified by a *reference*, a relatively small amount of information that can be used to access the object. A reference can be represented as a "pointer" or a "machine ad-

dress," which may require, for instance, only sixteen, thirty-two, or sixty-four bits of information, although there are other ways to represent a reference.

In some systems, which are usually known as "object oriented," objects may have associated methods, which are routines that can be invoked by reference to the object. An object may belong to a class, which is an organizational entity that may contain method code or other information shared by all objects belonging to that class. In the discussion that follows, though, the term object will not be limited to such structures; it will additionally include structures with which methods and classes are not associated.

Modern programs often run on systems using many processors and dynamically generate objects that are stored in a part of memory referred to in the field as the "heap." Although there are some different uses of the term, the discussion that follows will use heap to refer to shared memory managed by automatic garbage collection. The garbage collector has control of and/or direct access and/or knowledge of the addresses, classes, roots, and other such detailed information about all live objects created in the system.

After an object is no longer needed, it sometimes becomes necessary to reclaim the memory allocated to the object in order to prevent the system from running out of memory as more and more temporary objects fill the heap. Such memory reclaiming is referred to as "garbage collection," or GC. Known GC is well described by Richard Jones and Rafael Lins in their book, "Garbage Collection Algorithms for Automatic Dynamic Memory Management," published by John Wiley and Sons, 1996. This book is incorporated herein by reference. A brief description of known GC systems and techniques follows.

Garbage collectors operate by reclaiming space that is no longer "reachable." Statically allocated objects represented by a program's global variables are normally considered reachable throughout a program's life. Such objects are not ordinarily stored in the garbage collector's managed memory space, but they may contain references to dynamically allocated objects that are, and such dynamically allocated objects are considered reachable, too. Clearly, objects referred to in the execution threads' call

stack are reachable, as are the objects referred to by register contents. And an object referred to by any reachable object is also reachable.

The use of automatic garbage collectors is advantageous because, whereas a programmer working on a particular sequence of code can perform his task creditably 5 in most respects with only local knowledge of the application at any given time, memory allocation and reclamation require a global knowledge of the program. Specifically, a programmer dealing with a given sequence of code does tend to know whether some portion of memory is still in use by that sequence of code, but it is considerably more difficult for him to know what the rest of the application is doing with that memory. By tracing references from some conservative notion of a "root set," e.g., global 10 variables, registers, and the call stack, automatic garbage collectors obtain global knowledge in a methodical way. By using a garbage collector, the programmer is relieved of the need to worry about the application's global state and can concentrate on local-state issues, which are more manageable.

15 Garbage-collection mechanisms can be implemented in a wide range of combinations of hardware and/or software. As is true of most of the garbage-collection techniques described in the literature, the present invention makes use of and is applicable to most such systems.

To distinguish the part of the program that does "useful" work from that which 20 does the garbage collection, the term *mutator* is sometimes used in discussions of these effects; from the collector's point of view, what the mutator does is mutate active data structures' connectivity. Some garbage-collection approaches rely heavily on interleaving garbage-collection steps among mutator steps. In one type of garbage-collection approach, for instance, the mutator operation of writing a reference is followed 25 immediately by garbage-collector steps used to maintain a reference count in that object's header, and code for subsequent new-object allocation includes steps for finding space occupied by objects whose reference count has fallen to zero. Obviously, such an approach can slow mutator operation significantly.

Other, "stop-the-world" GC approaches use somewhat less interleaving. The 30 mutator still typically allocates space within the heap by invoking the garbage collector, for example, and the garbage collector, at some level, manages access to the heap. Basically, the mutator asks the garbage collector for a pointer to a heap region where it

can safely place the object's data. The garbage collector keeps track of the fact that the thus-allocated region is occupied, and it will refrain from allocating that region in response to any other request until it determines that the mutator no longer needs the region allocated to that object. In stop-the-world collectors, the task of memory reclamation is performed during separate garbage collection cycles. In such cycles the collector interrupts the mutator process, finds unreachable objects, and reclaims their memory space for reuse. As explained later when discussing "card tables," the GC's finding of unreachable objects is facilitated by the mutator recording where in memory changes have been made.

Garbage collectors vary as to which objects they consider reachable and unreachable. For the present discussion, though, an object will be considered "reachable" if it is referred to by a reference in a root. The root set includes, for instance, reference values stored in the mutator's threads' call stacks, the CPU registers, and global variables outside the garbage-collected heap. An object is also reachable if it is referred to by another reachable object. Objects that are not reachable can no longer affect the program, so it is safe to re-allocate the memory spaces that they occupy.

A typical approach to garbage collection is therefore to identify all reachable objects and reclaim any previously allocated memory that the reachable objects do not occupy. A typical garbage collector may identify reachable objects by tracing objects pointed to from a root, tracing objects pointed to from those reachable objects, and so on until all the referenced or pointed to objects are found and are retained. Thus the last objects found will have no pointers to other untraced objects. In this way unreachable objects are in effect discarded and their memory space becomes free for alternative use.

However, such free space is more useful when it is compacted than when it is distributed in a fragmented way throughout the heap. Compaction increases the data's "locality of reference." This increases cache hits and therefore cache performance. To compact free space, many garbage collectors may relocate reachable objects. In one known technique the heap is partitioned into two halves, hereafter called "semi-spaces." Between any two garbage-collection cycles, all objects are allocated in one semi-space ("from" space), leaving the other semi-space ("to" space) free. When the garbage-collection cycle occurs, objects identified as reachable are "evacuated," i.e.,

copied compactly into the "to" semi-space from the "from" semi-space, which is then considered free. Once the garbage-collection cycle has occurred, the designations "from" and "to" are interchanged for the next GC cycle. Any new objects will be allocated in the newly labeled "from" semi-space until the next GC cycle.

5 Although this relocation requires the extra steps of copying the reachable objects and updating references to them, it tends to be quite time and code efficient, since most new objects quickly become unreachable, so most of the current semi-space is actually garbage. That is, only a relatively few, reachable objects need to be relocated, after which the entire semi-space contains only garbage and can be pronounced free for 10 reallocation. One limitation of this technique is that half the memory so used is unusable for storing newly created objects.

A way of not only reducing collection-cycle length but also increasing overall efficiency is to segregate the heap into one or more parts, called generations, that are subject to different collection policies. New objects are allocated in a "young" generation, and older objects are promoted from younger generations to older or more "mature" generations. Collecting the younger generations more frequently than the others yields greater efficiency because the younger generations tend to accumulate garbage faster; newly allocated objects tend to "die," while older objects tend to "survive." But 15 generational collection greatly increases what is effectively the root set for a given generation since references to objects in one generation may be found in another generation, and thus other generations must be searched to uncover such references.

Consider Figs. 1 and 2, which depict a heap as organized into an old generation 14 and a young generation 16. With such a partition, the system may take advantage of a copy type GC's simplicity in managing the young generation because the unused half 20 memory is relatively small. But, for the "old" generation, which uses the great majority of the memory, using only half for storage may not be practical. So a different approach may be used. Among the possibilities are the mark-sweep and mark-compact described 25 in the above referenced book by Richard Jones and Rafael Lins.

In multiprocessor systems, one approach to speeding up garbage collections is 30 to "parallelize" the GC process by involving any idle processors in the garbage collection task. Toshido Endo et al. in their paper, "A Scalable Mark-Sweep Garbage Collector on Large Scale Shared-Memory Machines," published in the Proceedings of

High Performance Networking and Computing (SC97) in 1997 describes such a system. This approach includes copying the work to lockable auxiliary queues from which work may be stolen. A thread whose queue is empty looks for a queue with work, locks that queue and steals half the elements from the locked queue.

5

## SUMMARY OF THE INVENTION

The inventive system parallelizes the basic sequential GC operation on collection work tasks by employing a number of execution threads.

A separate work queue, with two ends wherein entries can be pushed and popped from the first end and at least popped from the second end, is created for each 10 thread. As a thread proceeds, it dynamically identifies referenced work tasks. Identifiers, usually pointers, are created for the referenced work tasks and pushed onto the first end of its work queue. At some later point, the thread will pop from the first end of its work queue the identifiers of the referenced work tasks, and in performing those tasks it may identify further work tasks. As before, identifiers of these further found 15 referenced work tasks are pushed onto the first end of its work queue for later processing.

When a thread has exhausted its own tasks (and any overflow tasks as described below) the thread steals work asynchronously from other threads' work queues by popping identifiers from the second ends thereof. By using the second opposite queue end 20 from the queue's "owner" thread, the stealing thread minimizes interrupting or blocking of the owner's operation.

Any contention that occurs is resolved, in a preferred embodiment of the invention, by using atomic instructions, i.e., instructions that represent operations that always run to completion without interruption by another instruction. As will be described 25 below, such instructions provide a way for the stealing thread to know whether it has been successful or not in stealing, and for obtaining a value useful for updating pointers when installing forwarding pointers to collected objects.

In one aspect, the present invention has adapted the work stealing to accommodate fixed size queues that can overflow. The queues' contents are popped and scanned

by their respective threads, and, if the scanning determines that there is not enough room available on a queue, an overflow list is created. In the case where work tasks are objects to be scanned for further referenced objects, such overflow objects are linked by class in this overflow list by replacing class pointers found in the objects with linking pointers.

### **BRIEF DESCRIPTION OF THE DRAWINGS**

Other objects, features and advantages will be apparent from the following detailed description of preferred embodiments thereof taken in conjunction with the accompanying drawings in which:

10 Fig. 1 is a circuit block diagram of a physical hardware system suitable for implementing the present invention;

Fig. 2 is a block diagram of the organization of the heap of Fig. 1;

Fig. 3 is a pictorial view of typical sources of roots of live objects;

15 Figs. 4a-4d are diagrams of the organization of queues, dynamic stealing, and linking of overflow pointers;

Figs. 5a-5d are block diagrams of the organization of the young heap;

Fig. 6 is a diagram of a card table heap relationship;

Fig. 7 is a flow chart depicting collection of the old generation;

Fig. 8 is a diagram of compaction in the old generation;

20 Figs. 9a-9c are diagrams of updating card table entries associated with old generation compactions, and

Fig. 10 is a block diagram of the termination organization.

### **DETAILED DESCRIPTION OF AN ILLUSTRATIVE EMBODIMENT**

25 Fig. 1 is a hardware block diagram of a multiprocessor computing system in which a number of processors 2 share common memory 8. Each processor may also have local RAM, cache, and mass storage, as well as communications and I/O devices. In addition there may be local communications paths 6 between the processors in the form of common busses, LAN's, or the like.

Fig. 1 depicts shared memory 8 as separate from memory disposed more locally to the individual processors, but this is not a necessary multiprocessor configuration. There may be a controller (not shown) for this shared memory in some embodiments. The operating system partitions the shared memory space into several sections, one 5 being the "heap." The heap is where space is allocated to objects dynamically under the control of a garbage collector. The heap itself is divided into two sections, referred to herein as the old generation 14 and the young generation 16. However, in other preferred embodiments, the heap may be divided into more than two sections. The old generation stores objects that have persisted (for some dynamically young generation 10 garbage collections), and the young generation stores newly created objects.

One or more operations in a GC cycle will normally be of the type in which the number of identifiable tasks to be performed expands dynamically. One such operation is that of identifying and relocating reachable objects, as a GC may do for, say, the 15 young generation. This operation starts with the tasks of identifying the objects to which root references point. These tasks are statically identified before the GC operation begins. But other tasks are identifiable by the GC only dynamically, i.e., in the course of performing previous tasks. In the course of performing the statically identified tasks, for instance, the GC threads will find reachable objects and in doing so 20 identify the tasks of finding objects referenced by those reachable objects.

In the illustrated embodiment, the collector divides initial, statically identifiable tasks into task groups. In the specific case in which the tasks are the processing of roots to find reachable objects, a convenient way to divide the tasks into groups is to group them in accordance with what their sources are. With reference to Fig. 3, typical 25 sources are: JAVA™ classes (class statics), stack frames, and native language interface roots (JNI). (JAVA is a trademark or registered trademark of Sun Microsystems, Inc. in the United States and other countries.)

When these tasks have been divided into groups, the groups are assigned, in a manner presently to be described, to respective GC threads. After a GC thread has 30 been assigned one or more such static task groups, it begins performing the tasks in its group or groups. By performing those tasks, the GC thread dynamically identifies further tasks to perform. In the specific case in which the operation is that of young-

generation reachable-object identification, the further tasks can be associated with respective reachable objects, because the tasks are following those objects' references to identify further reachable objects.

In accordance with the present invention, as shown in Figs. 4a, 4b, 4c and 4d, 5 each GC thread has a respective work queue 30 to keep track of these further tasks. As it dynamically identifies such tasks, the thread places into its work queue entries for each such task. In the case of the reachable-object-identification operation, for example, a convenient type of work queue entry is a pointer to the reachable object that the GC thread has found. That pointer will represent the task of following pointed-to objects' references to find further reachable objects. 10

Of course, other task granularities are possible. A separate entry could be made for each reference in a newly identified reachable object, for example.

As will be discussed further below, a GC thread performs the tasks in its work queue until that queue is empty, and it then searches other threads' queues for tasks to 15 steal and perform, as will be also explained in more detail. Before addressing that "work-stealing" feature, though, task-group assignment is discussed in more detail.

A competition is one preferred way of assigning groups to threads. That is, the GC threads can take turns claiming task groups that other threads have not yet claimed. For this purpose, the GC would maintain a flag associated with each task group to indicate whether that task group has been claimed. 20

A convenient way to synchronize task-group claiming is for the GC thread to use an atomic "test-and-set" operation. In an atomic test-and-set operation, a processor atomically reads an old value from a location and writes a new value into that location. For task-group claiming, the thread performs a "test-and-set" operation on the location 25 of a task group's flag, and the new value to which the flag is set is one that indicates that the task group has been claimed. If the flag's old value indicated that the task group had already been claimed, then the thread concludes that it was not successful in claiming the task group. Otherwise, it concludes that it was successful, and it will process that task group.

Preferably, there are more task groups than GC threads, e.g., four times as many task groups as threads, so some or all the threads claim more than one task group. Preferably each GC thread will claim a further task group only after it has finished processing the previously claimed task group. In any event, the processing often results in identification of new tasks to be performed, and a thread adds such tasks to its work queue by pushing an identifier of the newly identified task onto the bottom of its work queue.

Having a smaller number of GC threads contend for a larger number of task groups works initially to balance the work of each thread. But, as the threads process the tasks, some tasks will lead to identifying a large number of tasks while others will lead to identifying few or none. In the case where the tasks to be performed are those of following references to reachable objects, some objects may refer to no other objects, while others may refer to many. So the load could become unbalanced if no control measures were taken.

The present invention dynamically balances the work among the available threads. For that purpose, it employs what was referred to above as "work-stealing." The basic technique of dynamic work stealing queues is described in a paper by Niraj S. Arora et al., entitled "Thread Scheduling for Multiprogrammed Multiprocessors," from the Proceedings of the Tenth Annual ACM Symposium on Parallel Algorithms and Architectures, 1998. This paper is hereby incorporated herein by reference. In the present application Figs. 4a-4c illustrate the work stealing technique. As mentioned above, each thread places into its respective work queue entries representing tasks that it has newly identified. In the reachable-object-identification operation, in which the GC thread scans each newly identified object for references to further objects, one possible form for such an entry to take is that of the further-object-identifying reference itself. In the work queue, such a reference represents a task that may include steps such as scanning the further object for references, relocating the object, performing necessary reference updating, etc.

Referencing Fig. 4b, the GC thread pushes newly found references 38 onto one end, which is arbitrarily referred to as the bottom, of its work queue 30 for later proc-

essing. When the thread is ready to perform a task from its queue, it will pop a reference 36 from the bottom of the queue and perform the represented task.

Control of the queue is implemented, with reference to Fig. 4a, in a preferred embodiment, with an index 32 pointing to the next entry to be popped from the top of the queue and an index 34 pointing to the location where the next entry should be added to the bottom of the queue. In addition there is a tag 31 that, in this preferred embodiment, is used to control the operations of the queue as discussed below. The GC thread associated with the queue increments the bottom index when it pushes a task identifier onto its queue, and it decrements that index when it pops an identifier from it.

Now, the GC thread makes work queue entries only for objects that still need processing. If all of a newly identified reachable object's references refer only to objects that have already been processed, that object's processing results in no further work-queue entries. So a GC thread may reach a point at which its queue has no more entries. When this happens, the thread will "steal" work if it can from other GC threads' queues. Such stealing threads use the top index 32 and tag 31 of the queue to be stolen from because those threads pop entries from the top of the queue, rather than from the bottom, as the queue's "owner" thread does.

We now describe one way to implement such queues by reference to sample simplified code, with designation "popTop" (for the top-popping), "popBottom" (for bottom-popping), and "pushBottom" (for bottom-pushing). First we consider popTop, which stealing threads perform as shown in Fig. 4c. To understand the tag field's purpose, it helps first to consider in detail how the step of popping the queue from the top would be performed without the tag field. The following is illustrative sample code that performs popTop by a stealing thread.

25 Worktask\* popTop ()

```

{
  1   oldAge = age;
  2   localBot = bot;
  3   if (localBot <= oldAge.top)
  4     return NULL;

```

- 12 -

```
5      task = deq[oldAge.top];
6      newAge = oldAge;
7      newAge.top++;
8      cas(age, oldAge, newAge); /*atomic compare-and-swap*/
9      if(oldAge == newAge)
10     return task;
11     return NULL;
}
```

To "steal" from the top of another thread's work queue, the stealing thread first reads that queue's top index, as line 1 above indicates, to find where its top entry is. In the illustrated embodiment, the tag field and the top index are part of the same word, i.e., are in a group of bits ("age" in the sample code) that can be accessed in a single machine-level operation, such as a load, store, or an atomic compare-and-swap operation. (As was mentioned above, the tag would be read in the same operation, but this will be discussed below.) It then reads the thus-identified queue entry and reads the bottom index to make sure that the bottom index is not less than or the same as the top index, i.e., that the queue is not empty. As line 3 and 4 above indicate, the stealing thread will not pop the top of the queue if the queue is empty.

Otherwise the stealing thread reads the top-index identified queue entry. But the stealing thread does not immediately perform the task that the queue entry identifies. This is because, after it has read the top index, the stealing thread may be suspended after reading the location to which the top entry points but before it increments the top index to indicate that it has claimed the task. If so, a second stealing thread may pop the top entry in the interim, and the first stealing thread could then end up attempting to process an object that the second thread had already processed. So, before it actually performs the task, the stealing thread performs an atomic compare-and-swap operation, as line 8 indicates, in which it effectively pops the top queue entry by incrementing the top index 32 if that index's value is still the same (indicating that there was no second stealing thread pop) as the one the stealing thread used to read the top queue entry. As line 9 indicates, the storing operation is actually performed on the entire word including the top 31 and the tag 32 of Fig. 4a rather than only on the top field, for reasons that we will turn to below.

The operation of the CAS (compare and swap) instruction above is more easily understood by the following short, noted code:

```

newAge = CAS(newAge, oldAge, &dq->age) /* this checks the contents of
5      of
      dq->age. If it is equal to the value
oldAge, then dq- is set to newAge
      and oldAge is returned. */

If (newAge == oldAge) {
10    /*The CAS succeeded*/
} else {
    /* The CAS failed */
}

```

If the stealing thread thereby successfully pops the queue, i.e., if the absence of a top-index-value change enabled it to increment the top index, it proceeds to perform 15 the task from line 10 above that the top entry indicated. With reference to Fig. 4c, if thread T1 steals from the queue 42 of thread T2, the successfully stolen task may contain a reference 35 which is stored on the stealing threads queue 40 for later processing. If a top index field change did occur, though, another thread has presumably already popped the queue entry. So, line 11 returns "NULL" value and the first stealing thread 20 concludes that it has not popped the top entry successfully, and it neither increments the top index (as part of the compare-and-swap operation) nor performs the task that the queue entry represents.

Thus employing an atomic compare-and-swap operation protects the pop operation's integrity from interference by other stealing threads. Without more, though, 25 the top-popping operation would still be vulnerable to interference from the (bottom-popping) owner thread. To understand this, first consider how the owner thread pushes queue entries.

Unlike stealing threads, the owner thread pushes and later pops entries from the bottom of the queue. As the following sample code illustrates for pushBottom, pushing 30 a queue entry is simply a matter of reading the bottom index (bot), writing an entry identifier of a work task (wkt) into the location that the bottom entry indicates, and incrementing the bottom index:

```
void pushBottom(Worktask* wkt)
```

-14-

```

    {
1      localBot = bot;
2      deq[localBot] = wkt;
3      localBot++;
4      bot = localBot;
    }

```

Since a stealing thread will not pop an entry if the top index equals the bottom index (step 13 below), and since the owner thread is the only thread that pushes or pops entries at the bottom of the queue, there is no need to take special precautions against 10 interference by other, stealing GC threads.

As the following sample code indicates, popBottom - popping an entry from the bottom of the queue is more complicated than pushing an entry there.

Worktask\* popBottom()

```

    {
15     1      localBot = bot;
2      if (localBot == 0)
3          return NULL;
4      localBot--;
5      bot = localBot;
20     6      wkt = deq[localBot];
7      oldAge = age;
8      if (localBot > oldAge.top)
9          return wkt;
10     bot = 0;
25     11     newAge.top = 0;
12     newAge.tag = oldAge.tag + 1;
13     if (localBot == oldAge.top) {
14         cas(age, oldAge, newAge)
15         if (oldAge == newAge)
30     15         return wkt;
    }
16     age = newAge;
17     return NULL;

```

{}

To pop a queue entry, the owner thread first reads the bottom index bot, as in line 1 above. When the queue is in its initial, unpopulated state, the bottom index has an initial value, call it zero, that represents the start of the space allocated to queue entries. If the bottom index's value is zero, as in line 3 above indicates, then the owner thread concludes that its queue is empty and the routine indicates this by returning NULL. The thread in response thereto attempts to steal from another thread's queue. Otherwise, if the queue is not empty, the owner thread decrements the bottom index's value and pops the queue location to which the decremented bottom index points, as lines 4 to 6 above indicate.

But a stealing thread may have popped an entry from that location, because the bottom popping location is one location above the one at which bottom pushing occurs, so the stealing thread's comparison of top index does not prevent it from popping the same entry as the owner thread. The owner thread must check the top index's value before it performs the task represented by the entry that it has read. If the top index's value plus one is less than the bottom index's, the owner thread considers the pop operation a success and proceeds with the task.

If the top index's value plus one is not less than the bottom index's, then the owner thread has read the last queue entry, so it sets the bottom index to zero so that any new entry will be placed at the beginning of the queue space rather than unnecessarily continuing the queue contents' downward progression. It also sets the top index to zero, but two factors complicate the way in which it does so.

The first factor is that the owner thread needs to determine whether a stealing thread has popped the entry the owner has just read. If, as tested in line 13, the value that the owner thread read for the top index represents a location lower than the one represented by the bottom-index value 34 it read, then a stealing thread clearly has already popped the queue entry that the owner thread read, so the owner thread simply sets the top index to zero, as lines 11 and 16 indicate, and turns to stealing without performing the task that the entry represents. On the other hand, if the values the owner read for the top and bottom indexes were equal, then the queue entry it read may not

already have been popped by a stealing thread. If that entry has not been popped, the owner thread should perform the task that it represents.

To determine whether the entry has been popped, the owner thread performs line 14's atomic compare-and-swap operation, in which it sets the top index to zero if the value of age before the swap is still what it read previously, i.e., if no intervening popping of that last queue entry has occurred. If that compare-and-swap operation is successful, the owner thread performs the queue-entry-indicated task that line 15's return value represents. Otherwise, it still sets the top index to zero, as line 16 indicates, but it turns to stealing in response to line 17's NULL value: it does not perform the task whose identities it read from the queue.

The second complicating factor brings us to the reason for the tag field. As was just explained, the owner thread resets the indexes to point to the top of the queue space when the queue empties. Not only could this occur while a stealing thread is suspended in the act of stealing, but the owner could also bottom push the queue enough to fill this queue back up to where it was when the stealing thread's suspension began, and other stealing threads could restore the top index to the value originally read by the suspended thread. For example, when an empty queue has one element pushed onto its bottom, and then a steal attempt occurs, the stealing thread could read the top index and associated task, but the owner thread could then pop that task, process it, and push a new task. In this example, the top index was originally zero, so no "help" from other stealing threads is required to return it to its original value. If, as described above, the stealing queue tested for intervening accesses by comparing only the top index value with the one it used to read the queue entry, it would fail to detect the intervening accesses' occurrence, and it would perform the task unnecessarily and improperly.

Therefore, the tag field of the new age value with which age's controlled or replaced in line 14's or 16's step will have been incremented, as line 12 indicates.

To enable the stealing thread to guard against this, the owner thread increments the tag value when it resets the top index to zero. Then, when the stealing thread checks for intervening accesses, it compares that whole word with what it was when it read the queue entry. So, even if the top entry seems not to have changed, the change in the tag

value will alert the stealing thread to the intervening access, and it will not inappropriately perform the indicated task.

The foregoing sample code was simplified in, among other things, that it assumes an unlimited queue size. In some implementations, the queue size may be limited and subject to overflow. In order to avoid using more memory for an overflow list or table, the present invention creates an overflow data structure, used by all threads, that takes advantage of the class pointer in the object structures used by most object-oriented languages such as the JAVA™ programming language. As will be discussed presently, the collector uses that pointer to link overflowed objects by replacing the class pointers with pointers to other objects on the overflow list.

In operation, before references from a newly scanned object are pushed onto a work queue, the thread checks free space to see if pushing another object pointer onto its queue would overflow the queue. If so, with reference to Fig. 4d, the thread first obtains a lock on the overflow data structure. When a thread obtains the lock, only that thread has control of the overflow data structure. The particular lock mechanism, several of which are known in the art and can be used for this purpose, is not critical.

Still referring to Fig. 4d, when the thread has obtained the lock on the overflow structure it proceeds to remove half the identifiers from the bottom of the queue, one at a time, and place them on the overflow list 58, in a manner that will now be described. The overflow data structure is a table in which each entry includes a class identifier 58 and a class pointer 59 that points to a linked list of objects representing tasks in the overflow list. To add a task to the overflow list, the thread reads the class field of the object that represents that task, and it thereby determines the object's class. If the overflow data structure does contain an entry that represents that object's class, the thread adds the task at the head of the list. It does so by writing into the class pointer a pointer to the new task, and replaces the class pointer in the new task with the pointer task that was previously in the class pointer. So the result is that the class pointer points to the new task and the new task class pointer points to the task that was previously at the head of the list. The overflow objects are listed by class so that during retrieval the proper class pointer can be re-installed in each object's header.

In some embodiments, the overflow data structure may be efficiently represented by storing a class's list of overflow objects directly in the class data structure, and maintaining the set of classes with non-empty overflow lists as a linked list of class data structures, threaded through another field of the class data structure.

5 If the overflow data structure has no entry that represents the object's class, the thread adds such an entry to that structure and adds a corresponding pointer 59 to the object as the first instance on the overflow list of a task representing object of this class. The thread also NULLs the class pointer in the object's header. This NULL remains since the first task on the overflow list in a class becomes the last task when other tasks  
10 are added in front of the first task. The NULL is used to indicate the last task in an overflow list for a class.

When a thread has exhausted its queue, it obtains a lock on the overflow data structure, and retrieves a number of the objects at the beginning of the list. In doing so the thread must replace the class pointer to point to the tasks not retrieved to maintain  
15 the linked list. In one preferred embodiment, the thread may retrieve enough objects to fill half of its queue, but other amounts may be retrieved. The retrieved objects are pushed onto the bottom of the queue. If half a queue's volume is retrieved, and this queue were to overflow again, the retrieved objects will be in the top half of the queue and hence those retrieved objects will never be placed again on the overflow list.

20 In most applications overflow occurs rarely, so that the delay in handling overflows will not appreciably slow down the collection operation.

As referenced above, the young generation is split into two equal sized semi-spaces, marked "from" and "to," as illustrated in Fig. 5a. As shown in Fig. 5a, root 44 points to object B, and root 46 to object C in the "from" semi-space, and the task is to  
25 copy B and C and their references into the "to" semi-space.

Since the thread handling the copying of B and the thread handling the copying of C operate asynchronously, they will be contending for writing into the still-unused portion of the "to" space. That contention must be resolved. As discussed several places herein, atomic hardware primitive instructions (e.g., CAS) are used to update the  
30 pointer delimiting the bottom of this unused portion. However, contention for this

pointer might occur too frequently, and unacceptably slow the GC operation. In order to address this particular issue, with reference to Fig. 5a and 5b, GC threads allocate local allocation buffers (LABs) in "to" space. The allocation of LABs uses atomic hardware instructions, but only the allocating thread can allocate within and copy into a 5 LAB that it owns. So less expensive non-atomic instructions suffice for allocation within a LAB. In subsequent copying operations, thread t1 copies into LAB 50 until LAB 50 no longer has enough room left to receive the next object.

When this happens, the thread allocates another buffer and fills the remaining space in the old buffer with a "dead object", such as an array of integers, that benignly 10 occupies the remaining area. Such a benign object contains information in its header that allows any program that accesses the dead object to know where the next object, i.e., the first one in the next LAB, begins. This preserves the heap as a contiguous area. Thread t1 then allocates a fourth LAB in the same manner. In the interim, threads t2 and t3 may have allocated LABs 51 and 52, as shown.

15 Each LAB may be set up with individual copy pointers that are used similarly to the global one mentioned above, but without requiring the use of atomic primitive instructions, to control and organize the storing, without interference, of objects in that buffer. Each LAB is designed to be large enough to reduce contention for the copy 20 pointer but small enough to keep fragmentation of the young generation's storage space acceptably low. The actual sizes will depend on the applications involved. In some preferred embodiments, a single LAB may contain many megabytes.

Figs. 5b, 5c, and 5d show the before, intermediate and after, respectively, conditions as threads copy objects from a "from" memory into LABs in a "to" memory. The arrows in these Figs. are pointers. Thread t1 copies B to B' and updates the root 25 pointer 44, and, asynchronously, t2 copies C to C' and updates the root pointer 46, both operations as shown in Fig. 5c and 5d. With reference to Fig. 5b, thread t1 scans B, and t2 scans C, finding both with a reference 45 to D, and both speculatively allocating blocks D' and D'' within their respective LABs, as illustrated in Fig. 5c. So D must be copied and the proper pointers installed in B' and C'. In the sequential (non- 30 parallel) version of the copying collection algorithm, the object is copied when the first reference to it is scanned, and a "forwarding pointer" is installed in its old location so

that subsequent references to the object can be updated to the new location. It is possible to distinguish an object with an installed forwarding pointer from one without a forwarding pointer. In the scenario just described, threads t1 and t2 both read the word in D that would contain the forwarding pointer and observe that none is present. In order to resolve the contention between t1 and t2 for who is to copy D, both t1 and t2 attempt a CAS instruction to insert the forwarding pointer in D. The instruction CAS (addr, old, new) atomically compares the contents of "addr" with the value "old", and, if they agree, sets the contents of "addr" to "new". In this use, the "addr" is the address of the forwarding-pointer word of D; the "old" value is the non-forwarding pointer value both threads read, and the new values for threads t1 and t2 are the new addresses D' and D", respectively. Only one of these CAS instructions will succeed and only the thread performing the successful CAS will actually copy D. That thread copies D to its new location, and the other thread updates its references to that location. In this case, say that t1 does so first. Thread t1 copies D to D', updates the reference in B' 55 and leaves behind in the location of D a forwarding pointer. With reference to Figs. 5c and 5d, thread t2 is second to execute its CAS on D, but t2 finds that another thread has handled D since the newly written forwarding pointer in D will not match that in the CAS instruction executed by t2. However, t2's CAS operation returns the newly written forwarding points to D' and, so, t2 updates the reference pointer 60 in C' to point to D' and de-allocates the block D" 49 that was previously allocated for copying D by t2. Thread t1 finds a reference 57 in D1 to object E. and copies E to E' again leaving a forwarding pointer in location E and updating the pointer in D' to point to E' 59.

When the GC collects the young generation, the root set includes references in the old generation to young generation objects. One approach would be to inspect all references in the old generation at the beginning of every GC cycle, and this approach may be feasible for certain applications. But such an implementation is too slow for many applications. Another, faster approach includes "write barriers" in the mutator process, and a "remembered set" data structure. The remembered set records the locations of previously identified cross-generational references. A "write barrier" identifies and records possible modifications to references that have been made since the prior collection cycle. Only these modified references may contain newly created cross-generational references. The collections become faster because the remembered set and

the write barrier direct the collector only to those places in the older generation that may contain previously existing or newly created references to young-generation objects

One approach for implementing the above technique is to build what is known as a "card table." Fig. 6 depicts the old generation partitioned into physical segments called "cards," 72, 74, 76, 78, and 80. For each "card" there is a respective entry in the card table 70, where entries 71, 73, 75, 77, and 79 are associated with cards 72, 74, 76, 78, and 80, respectively. The card table entries are, in this embodiment, made by the write barrier code in the mutator. That code sets card table entries to identify their respective cards as possibly containing pointers to the young generation. One example of the information that may be found in the card table entries is: empty, meaning there are no young generation references in the respective old generation partition; modified, meaning that there may be references therein; or summarized, meaning that offsets to the references are contained therein. Fig. 6 shows that a reference 69 exists from object K in old generation card 72 to object A in the young generation. The possible existence of this reference would be indicated by the contents of card table entry 71. Since the organization of the old generation is fluid and the objects do not necessarily start and stop at the card boundaries, the entry in the card table contains an offset with respect into the start of the card. That offset indicates the starting point of an object that might straddle a card boundary. (Thus, the card table both embodies a remembered set and is used by the write barrier to track modified references.)

At the beginning of a collection cycle, the card table is scanned to find non-empty old-generation cards. Summarized cards contain young-generation references found in previous collection cycles. Modified cards may contain young-generation references created since the last collection; when these cards are examined, the corresponding card table entries are set either to "empty" or "summarized", depending on whether or not young-generation references were found. So that different processors can perform parts of this operation, the collector divides the card table into partitions, and each thread claims different partitions and finds references therein that point to the young generation.

In practice it has been found that old objects that reference young objects may tend to crowd into adjacent memory cards rendering those very densely populated with objects needing processing. For this reason, when a series of adjacent card table entries indicate that the corresponding adjacent old generation cards must be scanned for 5 young object references, the threads are arranged to scan an old generation card and then to skip a number of cards before scanning another. In one embodiment, partitions are formed of two cards separated by seven other cards. Consider that there are eight threads where each thread skips the seven intervening cards being handled by the other seven threads. In other preferred embodiments, the number skipped may be different or 10 even unrelated to the number of threads. By handling the cards in this card skipping or card striding manner, more threads will share in the processing of densely populated regions.

The following illustrates the advantage of the above process. When a card table entry indicates that the corresponding card must be scanned, the work task involved is 15 to find those objects in the old generation with references to objects in the young generation, and subsequently scan those young objects for further reachable young generation objects. Since these work tasks could be time consuming, the above parallelization works to share the work with all the available threads which speeds up the collection process.

20 The assignment of card partitions to threads is accomplished by competition among the threads. However, in other embodiments, the card partitions may be simply assigned. As discussed previously, an atomic set-and-test operation in claiming a card group resolves contention. A flag bit is associated with each card partition and a thread interrogating that bit with the test-and-set instruction will be successful or not and process 25 the card partition or go on to compete for another.

In a preferred embodiment, when an object survives a given number of young generation collection cycles, usually 3 or 4, that object is not copied to the "to" semi-space but is moved to the old generation. The GC maintains a pointer to free area(s) in the old heap and copies the object to one such location. As always, when an object is 30 moved, a forwarding pointer is installed in its old location, and that pointer is used to update all other references to the object to point to the correct new location.

The old generation will be subjected to a garbage collection when such an addition to the old generation will overflow the old generation, or, in other embodiments, other criteria may be used to trigger a GC cycle. GC in the old generation will occur substantially more rarely than in the young generation.

5 As mentioned above, for large heaps it may be inefficient to use the "from" and "to" semi-spaces since one is empty. The approach to old generation collection in this embodiment is based on the known "mark-compact" type collector, except the present invention adds parallelization to the mark-compact process. This operation includes the general steps, described in more detail below. With reference to Fig. 7, the steps include: marking the live objects; sweeping the heap to discover unmarked objects and count the live data; calculating where parallel compaction will move objects and installing the resultant forwarding pointers 128; updating pointers 130 to the live objects; and compacting the live objects and adjusting the associated card table 132, 134.

10

With reference to Fig. 7, item 120 the GC in parallel marks all the live objects  
15 in the heap by statically partitioning the root set into a number of segments usually larger than the number of threads. The segments are assigned to the threads. The threads scan the segments for object references. Referenced objects not already marked live are marked, and pointers thereto are pushed onto the thread's work queue, to indicate the pending task of scanning that object for references to other objects. After  
20 pushing an object reference on the queue, the thread may immediately pop and process the scanning task (which may in turn identify more such tasks, and so on), or it may continue scanning the root for object references. This choice might be made dynamically, depending on how full the thread's work queue is. When the scanning of the root for references is complete, then, the threads' tasks are to pop each succeeding object  
25 from their respective queues, to mark the live objects by setting a bit in the objects' headers, and to scan the object for other references. When other references are found, those objects are marked and scanned for yet further objects. The thread continues until all the objects and their references along a reference chain are marked. When a thread has exhausted its work queue of objects to be marked, the thread will steal from  
30 other threads, as discussed above with reference to Figs. 4a-4e.

Once all the live objects are marked, the heap is again partitioned 122 into a number of "units" greater than the number of GC threads, again to provide a parallelization of the GC. In one embodiment, the number of units is four times the number of threads, but a finer-grained partitioning into more units may be used to advantage. In 5 the fashion described before, each thread dynamically competes for and claims a unit to process by using atomic instructions accessing a flag. When a thread has control of a unit, it sweeps the unit coalescing consecutive unmarked objects into single free blocks, and records the number of bytes 124 occupied by live objects in the unit. When a 10 thread completes a unit, it goes on to compete for other units until all the units have been scanned.

Once the exact amount of live data is known, the heap is partitioned once more 126 into as many new "regions" as there are threads, wherein each new region contains substantially equal amounts of live data. Each region is composed of an integral number of units, whose live data occupancy was computed in the previous pass, allowing 15 the amount of live data in the regions to be balanced. This balancing is important because the region partition is used to parallelize the compaction phase, where cost depends on the amount of live data copied to new locations.

Before compaction, however, post-compaction addresses and corresponding 20 forwarding pointers are installed 128 and all object references are updated 130 so that they will be correct after compaction. Since the region boundaries are known, and the number and locations of the objects and the direction they are to be compacted within each region are also known, the threads can compute the address each object will have after compaction, and install a forwarding pointer to that address in each object before the compaction occurs.

25 With reference to Fig. 7 step 130, object pointers to old generation objects must be also redirected to the new addresses. Such pointers may occur in a number of different locations: in roots, in objects in other generations, and within objects in the current generation. These pointers occurring in items outside the current generation may be partitioned to equal the number of threads and each partition assigned to a GC 30 thread, and pointers in items within the current generation are partitioned using the over partitioning approach discussed above.

The last two operations in this sequence are to actually compact 132 the objects and then to update 134 the card table entries, item 70 of Fig. 6, to retain the indicators of old memory partitions having objects with references into the young generation.

Fig. 8 shows the resulting compaction of the old generation objects in a preferred embodiment where there are eight sections 80, 81, 82, 83, 84, 85, 86, and 87 each being compacted by the respective threads, T1, T2, T3, T4, T5, T6, T7, and T8. Each thread is assigned a partition to compact as shown. In compacting, alternate threads move the live objects in opposite directions. For example, the directions 88 and 90 of adjacent partition 80, 81 work to form a contiguous free area 92 adjacent to a contiguous filled area 100. In this way the old generation GC creates a series of spaced free regions 92, 94, 96, and 98 bordered by a series of contiguous filled regions 100, 102, 103, 104, and 105.

In Fig. 8, the compaction produces four separate free areas because the memory was divided into eight regions. The general rule is that the memory when divided into  $n$  regions, and when compacted as described above, will provide a number of contiguous free areas equal to the greatest integer less than or equal to  $(n+1)/2$ . So eight regions provide four free areas, and nine regions will provide five free areas. Although these free areas are not contiguous, in many applications, a small number of free areas are as useful as one large free area.

Since the old generation has been compacted as discussed above, the card table entries indicating those parts of the old generation having references into the young generation must be realigned with the newly compacted old generation. The technique is to "slide" the card entries as shown in Figs. 9a, 9b and 9c. This technique is illustrative and can be applied to memories that are organized and/or divided differently.

Fig. 9a shows a particular region 74 of Fig. 6 before it is compacted. For illustration purposes, the region is divided into four cards 110, 112, 114, and 116, with card table 73 having four corresponding card entries. Say that this region is to be compacted toward the left as shown in Fig. 9b, and as was done in regions 80, 82, 84, and 86 in Fig. 8. Referencing back to Fig. 9a, there are objects 140 that extend from the location 142 in card 114 to location 144 in card 116. The edges of the objects and the cards are not necessarily aligned. Also, there are objects 146 in card 110. In the ac-

companying card table 73, there is a "mod" entry 114' that indicates that the associated card 114 contains objects that have been modified, so one or more of those objects may have references into the young generation. The same circumstances are shown for card 116 and its card entry 116.' Card 112 is empty 112.' However, card table entry 110' 5 has "ref" meaning that one or more objects in card 110 have references into the young generation, and that nothing was changed in section 110. In the circumstances shown, the write barrier (discussed above) would not mark card 110 and the card table entry 110.'

When portion 74 is compacted to the left, across the empty card 112 and into 10 110, the associated card entries must also be moved left. The result is shown in Fig. 9b, where the objects 146 are moved to the left boundary 150 of card 110 and the objects 140 are moved left abutting 146. Here cards 114 and 116 are free space. The card table entries are modified where the effect is that the "mod" entries in 114' and 116' have 15 moved left by as many cards as the left boundary 142 has moved. Card table entries 114' and 116' are now marked as empty. Now when the card table is scanned for cards to search for entries into the young generation cards 110 and 112 will be searched.

If the region 74 was to be compacted to the right, as shown for regions 81, 83, 85, and 87 in Fig. 8, the operation is shown with the result shown in Fig. 9c. The right edge 144 is aligned with the right boundary 152 of card 116, and the left edge 142 re- 20 mains in card 114. Meanwhile, the object 146 is moved right abutting edge 142 as shown in Fig. 9c. Here card table entries 116' and 114' show "mod" indicators as before, and so the cards 116 and 114, respectively, will be searched (including 146) for references into the young generation, and, card table entries 110' and 112' will now be marked empty.

25 In the above operation, if the card table entries were summaries showing the addresses or offsets of live objects into the respective sub-regions, those addresses and offsets are known. Since, the compactor knows where the objects are and where they will finally reside; it can calculate new addresses and/or offsets into the proper card table entries.

30 The description so far has described a method of dividing among threads the various tasks dynamically identified during part of garbage collection cycle, and it has

given examples of garbage collection process parts that may use such a method. Since such a part of the cycle identifies tasks dynamically, there must be some way for a thread to determine when no further tasks remain in that part of the cycle. Fig. 10 illustrates one way to make that determination.

5 Fig. 10 shows four thread work queues, T1, T2, T3, and T4, and a global status word 140 containing a bit for each thread. If a queue is not empty, the corresponding status bit is set to active. This indicates that the associated thread has work remaining. If the thread runs out of work, it makes two times the number of threads attempts to steal. If unsuccessful, the thread sets its status bit to inactive, and then checks the status 10 bits of all the other threads. If all are set inactive, then the thread terminates the collection. If not all are inactive, the thread randomly selects a thread to steal from, and 15 peeks to see if that thread has work. If so, it sets its status bit active, and attempts to steal from that thread. If successful, the thread processes the work and starts the protocol over again. If unsuccessful, then the thread sets its status bit inactive and randomly chooses another thread to steal from.

Pseudo code for the termination procedure is as follows:

```

/* Try to find a deque with work and steal one item of work */
static java_lang_Object *stealWork(localDeque *dq) {
20    globalDeques *gdqs = dq->gdeques;
    int degree = gdqs->numDeques;
    int self = dq->index;
    int iterations = 2 * degree;
    int i = 0;
25
    while (i++ < iterations) {
        localDeque *dqToSteal = pickQueueToStealFrom(gdqs, dq);
        if (dqToSteal->bot > dqToSteal->age.top) {
            java_lang_Object *obj = (java_lang_Object *)popTop(dqToSteal);
30            if (!obj) poll(NULL, NULL, 0);
            else return obj;
        }
    }
    return NULL;
35 }

/* Look for work in the overflow list. If you don't find it,
   try to steal work from another thread */
static java_lang_Object *findWorkHelper(localDeque *dq) {

```